文章编号：1003-0077(2020)04-0030-08

中文症状知识库的建立与分析

咎红英[1,2]，韩杨超[1,2]，范亚鑫[1]，牛承志[2,3]，张坤丽[1,2]，穗志方[2,4]

(1. 郑州大学 信息工程学院，河南 郑州 450001；
2. 鹏城实验室，广东 深圳 518052；
3. 郑州大学第一附属医院，河南 郑州 450001；
4. 北京大学 计算语言学教育部重点实验室，北京 100871)

摘 要：构建大规模的知识库是人工智能、自然语言理解等领域的基础任务之一。症状作为描述病人的主观感受和诊断疾病的重要依据，更是优化智能导诊、医学问答等任务的重要因素。该文在现有的医学症状知识库研究的基础上，结合症状的概念、特征及在医学诊断中发挥的作用，构建了一个公开的中文症状知识库。该知识库从症状的本体分类、相关疾病、发作部位及多发人群等层面对相关属性进行了详细描述，涵盖了 8 772 种症状，共计 146 631 条属性关系。所构建的症状知识库(CSKB)是中文医学知识图谱的重要组成部分，并为 KBQA、知识推理及决策支持等应用提供了数据基础。

关键词：中文症状知识库；医学知识图谱；知识标注

中图分类号：TP391 **文献标识码**：A

Construction and Analysis of Symptom Knowledge Base in Chinese

ZAN Hongying[1,2], HAN Yangchao[1,2], FAN Yaxin[1], NIU Chengzhi[2,3],
ZHANG Kunli[1,2], SUI Zhifang[2,4]

(1. School of Information Engineering, Zhengzhou University, Zhengzhou, Henan 450001, China;
2. The Peng Cheng Laboratory, Shenzhen, Guangdong 518052, China;
3. The First Affiliated Hospital of Zhengzhou University, Zhengzhou, Henan 450001, China;
4. Key Laboratory of Computational Linguistics, Ministry of Education, Peking University, Beijing 100871, China)

Abstract: Building a large-scale knowledge base is an essential task in the fields of artificial intelligence and natural language understanding. As an important basis for describing the subjective feelings of patients and diagnosing diseases, symptoms are important factors in optimizing tasks such as intelligent consultation and medical question answering. Based on the existing researches, this paper constructs an open Chinese symptom knowledge base according to the concept and characteristics of symptoms and their roles in medical diagnosis. The knowledge base describes the relevant attributes such as ontology taxonomy of symptoms, related diseases, body parts stroke, and the suffering populations, covering a total of 146,631 attribute relationships of 8,772 symptoms. The constructed symptom knowledge base is an important part of the Chinese medical knowledge graph, providing a data foundation for applications such as KBQA, knowledge reasoning and supporting decision making.

Keywords: Chinese symptom knowledge base; medical knowledge graph; knowledge annotation

收稿日期：2019-09-16 **定稿日期**：2019-10-16

基金项目：国家社会科学基金(18ZDA315)；河南省高等学校重点科研项目(20A520038)；河南省科技攻关项目(192102210260)；河南省科技攻关计划国际合作项目(172102410065)；河南省医学科技攻关计划省部共建项目(SB201901021)

0　前言

症状(symptom)是指病人主观感受到不适或痛苦等异常感觉或某些客观病态改变，是临床诊断的主要依据，对于早期发现疾病、诊断疾病具有重要意义。了解病人各种症状的发生和演变，是临床工作中非常重要的内容，因为病人的感受是在其病理生理症状的表现基础上发生的，对疾病的反应是其他检查不能替代的。了解症状是医生向病人进行疾病调查的第一步，是问诊的主要内容。症状各种各样，同一疾病可能有不同的症状，不同的疾病有可能有相同的症状，若仅凭单一或几个症状就草率地做出诊断有可能会造成误诊，这对于辅助诊疗和智能导诊等自然语言处理应用都带来了极大的挑战。

从人工智能的概念提出开始，构建大规模的知识库(KBs，knowledge bases)一直都是人工智能、自然语言理解等领域的核心任务之一，是构建知识图谱的基础工作。袁凯琦等[1]提出，知识图谱是语义网和知识库的融合和升华。同时，构建一个大型的知识库也可以支持 KBQA[2]、知识推理及支持决策等自然语言处理任务。

知识库目前可以分为两种类型：Curated KBs[3]和 Extracted KBs[4]。Curated KBs 是从维基百科和 WordNet 等知识库中抽取大量的实体及实体关系合并为结构化的信息。Extracted KBs 则是直接从上亿个网页中抽取实体关系三元组。与 Freebase 相比，这样得到的知识更加具有多样性，而其实体关系和实体更多的则是自然语言的形式，精确度要低于 Curated KBs。目前行业内使用的比较多的还是 Curated KBs，主要是因为 Curated KBs 比较简单，容易构建且噪声少。

医学知识库是对医学领域信息的模型化表示。我国最早的医学知识库研究源自于 20 世纪 70 年代末的中医专家系统及 90 年代的医学数据库，其中的典型代表有《中国医院知识仓库》(CHKD)、《中国生物医学文献数据库》(CBMdisc)、《中国疾病知识总库》(CDD)以及中国科学院开发的大型中医药文献库——《中国中医药数据库》(该知识库是目前世界上最具权威的中医学知识库之一)。其中，中国疾病知识总库(China Disease Knowledge Total Database，CDD)是由解放军医学图书馆与重庆维普咨询有限公司合作研发的一个面向临床医药学专业人士，同时兼顾大众的知识型数据库[5]。《中国生物医学文献数据库》(CBMdisc)是中国医学科学院医学信息研究所开发研制的综合性医学文献数据库[6]。中国中医药文献数据库收录了 1984 年以来国内出版的生物医学及其他相关科技期刊的中医药文献近 59 万余篇，覆盖中医药学、针灸、气功、按摩、保健等方面的内容。中国中医科学院中医药信息研究所[7]自 1984 年开始进行中医药学大型数据库的建设，目前数据库总数 48 个，数据总量 120 余万条，包括中医药期刊文献数据库、疾病诊疗数据库、各类中药数据库、各类国家标准数据库(中医证候治则疾病、药物、方剂)等相关数据库。

近年来，随着大数据与人工智能技术在医学领域的广泛应用，研究的主题也逐渐向多元化发展，包括电子病历、临床路径等在内的各类专科专病知识库建设成为研究热点，如要芳[8]基于本体的电子病历知识库的研究，实现了电子病历的分析本体和语义检索功能，郑西川等[9]在临床路径的知识库研究中，提出了知识库表达模型的三个组件，建立了知识库的规范化表达，用以支持病人的诊疗计划，Marek 等[10]使用文本处理和示例学习建立了疾病症状模型。

尽管国内在医学知识库方面的研究已初有成效，但从整体情况看仍处于各自为政的状态，缺乏深度，且对于症状知识的扩展分类和处理目前还没有相对统一的标准。而且，在已经提出的知识库中，关于症状的发作部位、多发人群及症状本体等重要分类信息上都存在着缺失、模糊的现象，难以直接应用于自然语言处理的研究。

华东理工大学在 OpenKG① 上发布了中文症状三元组关系库，这是一个包含症状实体的三元组的数据集。中文症状库的数据来自 8 个主流的健康咨询网站、3 个中文百科网站和电子病历。其还包含了中文症状与 UMLS 中概念的链接结果。该数据集还提供了关键词查询和 SPARQL 查询功能。但该知识库也有着一些不足之处，例如，三元组的描述体系不够完整、错别字多以及实体名格式不规范等，以至于该知识库的质量并不能作为标准支撑基于知识库的各种应用。

综上所述，本文为构建一个全面的知识库，从症状的三元组信息出发，从 OpenKG、CMeKG(中文医学知识图谱，Chinese Medical Knowledge Graph)②、电

① OpenKG 网址为：http://www.openkg.cn/

② CMeKG2.0 网址为：http://cmekg.pcl.ac.cn/

子病历、医学教材和其他主流的医学综合网站等媒介收集和整理了以症状为核心的中文症状知识库(CSKB①,Chinese Symptom Knowledge Base)。本文主要介绍了症状知识库的构建过程,总结症状信息在标注过程中遇到的典型问题,并提出了相应的解决方案。

1 症状知识库相关研究

由于目前国内在建立中文医疗病症知识库方面仍处于起步阶段,用于研究的公开医学知识库也不是很多,为了获得准确全面的症状信息,本文首先考察了国内外关于医疗知识资源库和医学知识图谱的构建现状,从医学教材的症状分类体系出发,主要用了以下三类主要资源构建症状知识库。

1.1 中文医学知识图谱 CMeKG

CMeKG[11]是利用自然语言处理与文本挖掘技术,基于大规模医学文本数据,以人机结合的方式研发的中文医学知识图谱。CMeKG 的构建参考了 ICD-10[12]、ATC、SNOMED、Mesh[13]等权威的国际医学标准以及规模庞大、多源异构的临床指南、行业标准、诊疗规范与医学百科等医学文本信息,通过对各类医学资源的整合,形成了多来源医学文本。借鉴已有的分类方法,对文本中的信息进行了抽取、整理与标注。对于这些非结构/半结构化的文本数据,采用了人工标注加自动提取两种方法从中提取关系,其中自动提取使用了规则加 tagging 模型的方法。对于抽取出的关系进行人工审核评估,从而构建出内容丰富的中文症状知识库。CMeKG 于 2019 年 1 月 30 日发布 1.0 版本,并于同年 8 月 1 日更新至 2.0 版本。

CMeKG 1.0 包括:6 310 种疾病、19 853 种药物(西药、中成药、中草药)、1 237 种诊疗技术及设备的结构化知识描述,涵盖疾病的临床症状、发病部位、药物治疗、手术治疗、鉴别诊断、影像学检查、高危因素、传播途径、多发群体、就诊科室等,以及药物的成分、适应症、用法用量、有效期、禁忌证等 30 余种常见关系类型,关联到的医学实体达 20 余万,CMeKG 目前的概念关系实例及属性三元组达 100 余万。CMeKG 的目标是建立大规模、高质量的中文医学知识图谱,为智慧医疗奠定专业知识基础。

CMeKG 2.0 在 CMeKG 1.0 的基础上进行了多维度,多层次的扩展与深化,对多源异构的医学资源进行人机交互的知识提取与知识融合,在此基础上增加了症状类知识,并对儿科疾病进行详细描述,CMeKG 2.0 目前已包含 11 076 种疾病,18 471 种药物,14 794 种症状,3 546 种诊疗技术的结构化知识描述,描述医学知识的概念关系实例及属性三元组数目达 1 566 494。与 CMeKG1.0 相比,CMeKG2.0 扩大了医学知识的覆盖面,进一步提高了其描述信息的丰富程度。

本文爬取了 CMeKG 中的疾病常见症状、术后常见症状、药物治疗后症状等,涉及 11 076 种疾病的症状信息共计 102 264 条。CMeKG 示例如图 1 所示。

	A	C	
1	entity1	sub_predicate	entity2
1081	利斯特菌脑膜炎	常见症状	严重的头痛
1082	利斯特菌脑膜炎	常见症状	伴意识障碍
1084	利斯特菌脑膜炎	多发群体	免疫功能缺陷成人
1085	利斯特菌脑膜炎	常见病因	利斯特菌感染
1088	利斯特菌脑膜炎	常见症状	可发生抽搐
1091	利斯特菌脑膜炎	常见症状	呕吐；脑膜刺激征明显
1094	利斯特菌脑膜炎	常见症状	如木僵
1095	利斯特菌脑膜炎	多发群体	婴幼儿
1108	利斯特菌脑膜炎	就诊科室	神经内科
1109	利斯特菌脑膜炎	多发群体	老年人
1112	利斯特菌脑膜炎	常见症状	谵妄等

图 1 CMeKG 示例

从图 1 可以看出,原始的 CMeKG 样例以疾病为中心,将症状作为疾病的一个关系组成的三元组,但由于该语料中的关系由机器+人工共同标注而成,所以原症状信息仍然存在错误,比如“如木僵”、“谵妄等”就属于在症状信息边界的切分上做得不是很好的例子。类似的错误还包括如“或刺痛”、“或剧烈绞痛”、“鼻孔朝天或”等停用词在前后出现的情况,可以通过构建停用词表批量处理相关信息。

1.2 OpenKG 中文症状库

原症状库来自于华东理工大学在 OpenKG 上发布的中文症状三元组关系库。本文摘取了该症状库中的 617 499 个三元组,共包含 135 485 个实体和 8 种三元组关系类型。OpenKG 中文症状库如图 2 所示。

从图 2 我们可以看出原始症状库的一些问题:

① 分类方式过于冗余,且原症状库中疾病、症状、科室、检查、药品等实体两两之间都通过“相关”

① CSKB 症状库资源访问地址为:http://www5.zzu.edu.cn/nlp/

	A	B	C
1	entity1	predicate	entity2
4	“鸡盲”或“雀盲”	症状相关疾病	雀蒙眼
6	β-氨基酸尿	症状相关症状	少尿
8	白色黏液状尿	症状相关症状	尿痛
9	遗尿肌无反射	症状相关科室	泌尿外科
10	比目鱼肌肌力异常	症状相关科室	神经外科
12	闭合容积(CV)	检查相关症状	痰咳
13	哺乳	疾病相关症状	涨奶
14	藏毛窦,藏毛病	症状相关药品	排石通淋口服液
15	藏毛窦,藏毛病	症状相关科室	眼科
16	层板状出汗不良	症状相关症状	干涸后残留环状鳞屑
17	出血性疾病	疾病相关症状	齿龈出血
22	大便脓血	症状相关科室	肛肠科
23	氮质血症	症状相关药品	包醛氧淀粉胶囊(析清)
26	低温性昏迷	症状相关疾病	遗传性血色病

图 2　OpenKG 中文症状库样例

来建立关系，但“相关”的定义表述不清，由此定义出的关系类型并不具有可信度。

② 由症状确定治疗药物本身就不符合医学常识，单依据症状和体征都不能决定治疗方法[14]。例如，头疼一症，病因有风、寒、暑、湿、火、血瘀、气滞等，性质又有寒、热、虚、实的差异，因此，仅凭“头疼”这个症状是不能决定治疗方法的，只有明白疾病的原因、部位、性质以及致病因素和病人的抗病能力的综合信息，才能有效地指导治疗的方法和药物用量。对于这种不可信的三元组关系，我们选择抛弃“症状相关药品”这类关系的词条。

③ 词条中有大量的病史、药品、手术、抗体、细菌、病毒、穴道、书名、部位等诸多与关系信息不对应的错误三元组关系，本文以规则的方法进行了错误过滤。

1.3　主流的医学综合网站

本文从垂直医疗网站 39 健康网①和寻医问药网②中获得以症状为核心的结构化数据，并存储在 MongoDB 数据库中。相关数据主要是以疾病为中心。数据形式包括结构化和非结构化。经统计，共从寻医问药网上爬取症状相关信息 8 765 条，从 39 健康网上爬取症状相关信息 8 225 条，其中的症状信息包括症状定义、症状起因、相关检查和相关治疗等。人工抽样校对后验证了该数据中的症状信息是规范化的语言，将校验过的数据存储在数据库中。

2　症状知识库的构建

根据已收集到的症状资源进行抽样检查后发现，三元组中所出现的问题主要分为三类：实体名错误、停用词与特殊符号未切分、实体关系的描述体系不统一。我们采取的处理方法如下：

① 对于多个知识源的症状信息取一个交集和差集，因为错误的信息较少同时存在于不同的知识源中。而对于差集中的症状知识信息，总结错误类型，如手术一般以"术"结尾，书名一般以“》”结尾，抗体药物中含有“抗”，药品名字中大量出现“酮”“醚”“素”等关键字样，病史则一般以“史”结尾等等，然后总结正则规律匹配字符串，将符合特征的词条提取出来，再进行人工校对，因为即使符合规律也有可能没有问题。（如“抗 β2-糖蛋白 1 抗体”不是症状，但“甲状腺过氧化物酶抗体过高”就是一种症状）。

② 其中的边界词以“或”“等”“如”居多。对于这种边界切分不清的情况，我们采取了统一的处理方法，即总结原语料中的常见边界词，对词条首尾正则匹配边界词后进行字符串替代为空格，使其变成可用词条，而对词条中间出现的边界词不予处理，尽量不改变词条本身原义。由于原语料的来源较为广泛，实体中还存在有特殊符号，采取的方法是将带数字、字母、特殊符号的字符串全部提取，逐一人工校对。

③ 在借鉴中文电子病历标注规范[15]的经验上，本研究初步制定出了关于症状知识描述体系，数据库的键值对共分为十个大类、症状名、定义、同义词、发作部位、所属科室、多发群体、相关疾病、常用检查、鉴别诊断及预防治疗。由于本数据集面向知识图谱的构建和问答，所以所给数据为三元组形式。

目前症状库共整理 8 772 种症状信息，涵盖了不同的＜症状—关系—属性＞的三元组共计 146 631 条。部分知识信息如下：{symptom：眼球内陷，definition：[眼球位置后退称为眼球内陷。是与眼球突出相反的状态，与眼球突出相比，较为少见。眼球内陷指的是由于眼球以外的原因所致的眼球内陷，必须与眼球缩小病变（小眼球、眼球萎缩、眼球痨）和随同发生的睑裂缩小所引起的眼球位置后退的假像区别开来。]，location：[眼｜头部]，department：[眼科]，population：[老人]，disease：[A-V 综合征，上睑下垂，小儿颈交感神经麻痹综合征，支气管肺癌，眼眶内静脉曲张，进行性面偏侧萎缩症，霍纳综合征]，…}

① 39 健康网网址：http://www.39.net/

② 寻医问药网网址：http://www.xywy.com/

3 分类标准

考虑到可能面对的应用场景，本文对已经收集好的共 8 772 种中文症状共计 146 631 条的三元组信息，从三种角度进行了描述：症状本体类型、发作部位及多发群体。症状信息在标注时可能会存在多个属性，在标注时就标出多个字符串属性值。

3.1 症状本体类型

本文对症状本体的分类进行了考察，给出了症状本体类型分类的描述体系。

将症状分为症状、体征、组合症状、疾病症状四种类型，并制定了同义词词表。其中，“症状”是病人自己向医生陈述(或是别人代述)的具体表现，如头疼、腹痛、鼻塞、恶心、呕吐等；“体征”是医生给病人检查时发现的具有诊断意义的征候，例如，白细胞增多、低血糖晕厥、费希尔试验阳性等；“组合症状”指当多个症状有紧密的伴发和并发关系时，形成的有规则性的知识信息，如咯血伴呛咳、右下腹痛伴呕吐等。具体信息如表 1 所示。

表 1 症状本体类型分类表

本体类型	症状数量	例子
症状	5 856	发热、心慌气短
体征	2 195	低血钾、逼尿肌反射亢进
组合症状	333	腹痛伴便血、关节间隙不规则和关节腔狭窄
疾病症状	388	牙龈出血、生长发育迟缓
同义词关系	77	构音困难现象/构音障碍、肘关节疼痛/肘痛

3.2 发作部位

目前已有的研究工作对症状的分类并不统一，根据发作部位的不同，本文将症状信息分为：头部(头、眉、眼、鼻、耳、口、牙)、颈部、胸部、背部、四肢(上肢、下肢)、腹部、盆腔、腰部、臀部、生殖部位(男性、女性)、皮肤、全身、其他(神经、精神、行为及血液循环系统等方面的症状)13 个大类，其中头部、四肢、生殖部位又根据细节的不同分为了若干个子类，具体分类信息如表 2 所示。

表 2 发作部位—症状分类表

类别	子类别	数量	例子
头部	头	514	头枕部慢性疼痛、颅内头痛
	眉	9	眉弓疼痛、眉毛外侧脱落稀疏而细
	眼	430	睑结膜泪阜区灰黑色肿物、兔眼
	鼻	151	鼻翼扇动、单侧鼻腔阻塞
	耳	115	耳前庭听觉失衡、复听
	口	342	口腔粘膜弥漫充血、舌苔发黑
	牙	70	牙齿咬合无力、牙龈肿胀
颈部		200	甲状腺肿大、颈根部斜方肌及风池穴处有压痛
胸部		696	胸闷、两肺弥漫或散在可逆性哮鸣音
背部		73	脊柱前突变直局部叩痛、驼背
四肢	上肢	324	肘部的肿胀压痛、屈腕屈指无力
	下肢	305	膝关节间隙弹响和疼痛、扁平足
腹部		713	胃痉挛、下腹剧痛并渐向腹中线扩散
盆腔		173	髋部酸胀不适、盆腔积液
腰部		222	持续性腰痛伴有晨僵、腰大肌痉挛
臀部		88	排粪时肛门灼痛、直肠疼痛
生殖部位	男性	138	睾丸发育不全、阴囊红肿
	女性	359	宫颈前唇水肿、阴道黏膜膨胀

续表

类别	子类别	数量	例子
皮肤		472	毛孔角化过度、猩红热样皮疹
全身		1 401	躯体或颜面某一局部的连续性抽动、疲倦乏力
其他(神经,精神,行为及血液循环系统等)		1 587	触觉等刺激反应过度、先天性铜代谢障碍

3.3　多发群体

部分症状会有多发人群的属性,根据普适性的分类方法,将多发人群分为老人、小孩、男性和女性四种类别,具体信息如表 3 所示。

表 3　多发群体-症状分类表

多发群体	数量	例子
老人	92	双手震颤、日常生活能力减退和行为异常
儿童	291	新生儿手足徐动、小儿夜啼
男性	91	前列腺肥大、附睾肿大发硬
女性	369	产后下腹坠痛或阴道坠胀感、淤血性乳房痛

4　结果分析

4.1　类别分析

按照以上规范和流程,历经一年时间,目前共收录、标注并人工校对的症状信息 8 772 种,对知识库中症状的类别进行统计得到的比例如图 3、图 4 所示。

从图 3、图 4 可以看出,发作部位在头部和全身的症状信息在种类上占据了很大的比例,而背部和臀部主要是由肌肉组成的,生病时产生病理反应的种类比较单一,从而占据的比例最小,且远小于平均值;从多发群体的类别分布上面可以看出,女性和儿童占据了很大的比例,通过观察语料发现女性为多发群体的症状大多与生产和育儿有关;儿童则由于脏腑娇弱、自身免疫系统未发育完全,所以在占据比例上这两者高于男性和老人。

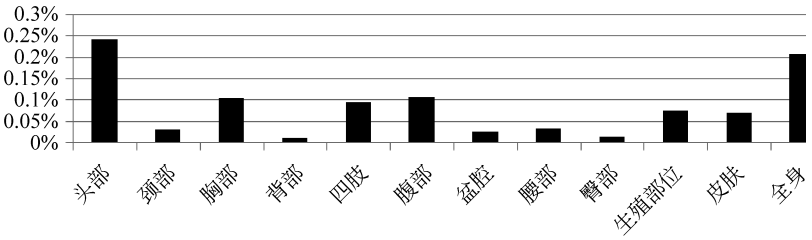

图 3　发作部位各类别比例

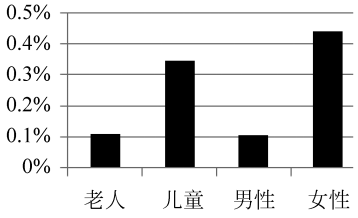

图 4　多发群体各类别比例

4.2　人工标注结果分析

为了检验知识库三元组的准确性,在医学专家的指导下,针对未经标注的多来源医疗文本进行了症状及症状关系的标注工作,共标注出有关 109 种常见疾病的症状的三元组关系 102 281 个。形式为<实体 1—实体 2—关系—属性>

标注样例:“食物中毒@中度脱水表现为口渴、躁动不安或易激惹的行为、皮肤弹性变差和凹眼。食物中毒@严重脱水时,症状更严重,可表现为休克、意识水平下降、少尿、四肢湿冷、脉搏细速、低血压或血压测不出、皮肤苍白。食物中毒@重度脱水患者还可能出现体位性低血压。”<口渴|躁动不安|……—食物中毒—症状相关疾病—中度脱水><休克|意识水平下降|……食物中毒—症状相关疾病—严重脱水><四肢湿冷—症状相关部位—四肢><皮肤苍白—症状相关部位—皮肤>

标注一致性一般可以用 Kappa 值[16]和 F_1 值[17]来表示。Kappa 值在情感极性分类的语料标注中应用较广，但在实体识别中，若把未标注的文本作为反例的话，则反例数量巨大难以统计，故可使用 F_1 值来对实体识别标注语料进行一致性评价，具体做法是将三标者的标注结果 B 作为标准答案，计算一标者的标注结果 A 的精确度(P)和召回率(R)，进而计算 F_1 值，计算如式(1)～(3)所示。

$$P=\frac{A_1 \text{ 和 } A_2 \text{ 一致的标注结果总数}}{A_2 \text{ 的标注总数}} \quad (1)$$

$$R=\frac{A_1 \text{ 和 } A_2 \text{ 一致的标注结果总数}}{A_1 \text{ 的标注总数}} \quad (2)$$

$$F_1=\frac{2\times P\times R}{P+R} \quad (3)$$

文献[18]指出，标注一致性达到 0.8 以上时，可以认为语料的一致性是可信赖的。最终，本研究标注的命名实体识别一致性达到了 0.873，实体关系一致性达到了 0.829，说明本研究的症状知识库的三元组是可信赖的。

5 应用及展望

将收集到的症状知识应用于医疗文本的分词系统，对处理医疗文本分析将会有帮助。症状的类别属性可以作为特征，加入到包括知识库问答在内的各种医疗文本的自然语言处理应用中；发作部位和相关科室属性对智能导诊很有帮助；问诊内容属性对辅助诊疗有很大帮助，属性信息可随症状实体本身一起用来辅助分词，也可以作为特征加入到机器学习应用中；也可为知识库问答提供帮助，一般知识库的问答系统方法都是基于语义解析的方法(SP-based)和基于信息检索的方法(IR-based)[19]，所以覆盖度更高的知识库会对问答产生更好的结果，也可为其他医疗语言资源建设提供扩展和帮助。

6 结语

本文主要对多来源的中文症状信息进行了总结和分类，初步构建了一个中文症状知识库(CSKB)。首先调研了国内外各类医学知识库资源的整体情况，将已有知识库的内容进行数据清洗后规范化，形成初步的语料库。与此同时，整合多来源医疗文本作为标注文本，和医学专家讨论制定出症状属性的描述体系，最后通过多轮迭代的方式标注医疗文本，并请医疗专家全程把控标注质量，以确保准确性，同时使用规则加机器学习的方法进行自动抽取，并对构建过程中的问题的解决方法进行了分析和解决。下一步，我们将尝试在包含医疗文本的自然语言处理应用中加入已经收集到的知识；同时以现有的数据作为基础，尝试利用半监督的信息抽取等技术进行有关医学知识的自动收录，减少人工收录、标注、整理的工作量，有效扩充症状医疗知识库的规模。

参考文献

[1] 袁凯琦，邓扬，陈道源，等. 医学知识图谱构建技术与研究进展[J/OL]. 计算机应用研究，2018，35(7)：15-22[2017-08-17]. http://www.arocmag.com/article/02-2018-07-068.html.

[2] Zheng H T，Fu Z Y，Chen J Y，et al. Novel knowledge-based system with relation detection and textual evidence for question answering research.[J].PloS one，2018，13(10)：e0205097.

[3] Yin P，Duan N，Kao B，et al. Answering questions with complex semantic constraints on open knowledge bases[C]//Proceedings of the 24th ACM International Conference on Information and Knowledge Management.2015：1301-1310.

[4] Downey D，Bhagavatula C S，Yates A. Using natural language to integrate，evaluate，and optimize extracted knowledge bases[C]//Proceedings of the 2013 Workshop on Automated Knowledge Base Construction. ACM，2013：61-66.

[5] 杨志. 一种基于知识挖掘与知识组织的知识型数据库——中国疾病知识总库之临床医药学知识服务系统介绍[J]. 中华医学图书情报杂志，2008，17(3)：63-65.

[6] 林亚君. 中国生物医学文献数据库与中国学术期刊全文数据库的比较[J]. 中华医学图书馆杂志，2001，10(3)：45-45.

[7] 刘红，华强. 中国中医药数据库网上检索系统用户使用情况分析[J]. 中国中医药信息杂志，2006，13(7)：98-99.

[8] 要芳. 基于本体的电子病历知识库研究[D].西安：西安电子科技大学硕士学位论文，2009.

[9] 郑西川，于广军，谭申生. 临床路径电子化与临床知识库建设实践[J]. 中国医院，2012，16(2)：29-31.

[10] Jaszuk M，Szostek G，Walczak A，et al. Building a model of disease symptoms using text processing and learning from examples[C]//Proceedings of the 2011 FedCSIS.M：187-194.

[11] 奥德玛，杨云飞，穗志方，等. 中文医学知识图谱 CMeKG 构建初探[J]. 中文信息学报，2019，33(10)：1-7.

[12] Sundararajan V，Henderson T，Perry C，et al. New ICD-10 version of the Charlson comorbidity index predicted in-hospital mortality[J]. Journal of Clinical Epidemiology，2004，57(12)：0-1294.

[13] Lipscomb C E. Medical Subject Headings (MeSH). [J]. Bull Med Libr Assoc，2000，88(3)：265-266.

[14] 吴其夏. 新编病理生理学[M]. 北京：中国协和医科大学出版社，1999.

[15] 杨锦锋，于秋滨，关毅，等. 电子病历命名实体识别和实体关系抽取研究综述[J]. 自动化学报，2014，40(8)：1537-1562.

[16] Carletta J. Assessing agreement on classification tasks：the kappa statistic[J]. Computational Linguistics，1996，22(2)：249-254.

[17] Hripcsak G，Rothschild A S. Agreement，the f-measure，and reliability in information retrieval.[J]. J Am Med Inform Assoc，2005，12(3)：296-298.

[18] Artstein R，Poesio M. Inter-coder agreement for computational linguistics[J]. Computational Linguistics，2008，34(4)：555-596.

[19] Fengyu Yang，Liang Gan，Aiping Li，et al. Combining deep learning with information retrieval for question answering[C]//Proceedings of the ICCPOL 2016 and NLPCC 2016. 2016：917-925.

訾红英(1966—)，博士，教授，主要研究领域为自然语言处理。
E-mail：iehyzan@zzu.edu.cn

范亚鑫(1997—)，学士，主要研究领域为自然语言处理。
E-mail：3262933823@qq.com

韩杨超(1994—)，通信作者，硕士研究生，主要研究领域为自然语言处理。
E-mail：841153012@qq.com